BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35433
doi: 10.3897/biss.3.35433

OPEN ACCESS

Conference Abstract

Automatizing the Detection of Erroneous Species Occurrence Records

Raul G Jimenez Rosenberg‡, Raul Sierra-Alcocer‡

‡ Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), Mexico City, Mexico

Corresponding author: Raul Sierra-Alcocer (raul.sierra@conabio.gob.mx)

Received: 14 Apr 2019 | Published: 21 Jun 2019

Citation: Jimenez Rosenberg R, Sierra-Alcocer R (2019) Automatizing the Detection of Erroneous Species Occurrence Records. Biodiversity Information Science and Standards 3: e35433.
https://doi.org/10.3897/biss.3.35433

Abstract

The work involved in checking millions of records by hand is hard and requires thousands of human hours. At the increasing rate at which we are collecting new data from different sources with a wide range of 'quality', the problem is getting worse. An institution like CONABIO (National Commission for the Knowledge and Use of Biodiversity, Mexico) dedicates a large amount of human resources to review species records to ensure that data published by the institution has high quality. At CONABIO we are designing a system to help us direct our attention to the most problematic data.

Our methodology (Stephens et al. 2019) scores a species record according to the features of its location, and it labels it as suspicious if it has a low score. A low score means that the features of the location are unusual for that species. The features of locations are the set of abiotic, like climate and topographic charactersitics and occurrences of other species in the location. Although this does not mean that a record is wrong, it may be an indicator that a record needs to be assessed.

The system we are designing works in two scenarios: in one, it scores new data based on parameters adjusted from validated data; in the second, the system checks for consistency in the database, that is, it flags records of a species that seem like outliers according to the predominant records distribution for that species. Our initial tests show that we could speed up the detection process for some problematic records. In one of our tests, where we used

data that were previously labeled by hand, the method flagged 624 records, out of which 70 were confirmed as incorrect data. If we look only at the precision of the results it might seem like a poor performance, however if we look at the amount of work it might save us, it looks promising because to find the same number of inaccurate records without any assistance we would have had to review almost 5,000 records.

This talk is a proof of concept for this system, and details on our initial results, reviewing both weaknesses and strengths.

Keywords

data cleaning, species occurrence records, statistical tools

Presenting author

Raul Sierra-Alcocer

References

- Stephens C, Sierra-Alcocer R, González-Salazar C, Barrios J, Salazar Carrillo JC, Robredo Ezquivelzeta E, del Callejo Canal E (2019) SPECIES: A platform for the exploration of ecological data. Ecology and Evolution 9 (4): 1638-1653. https://doi.org/10.1002/ece3.4800